30.039 Theory and Practice of Deep Learning
Project Proposal

Victoria Elizabeth Yong Shu Qi 1004455 | Vincent Leonardo 1004544
Dharmapuri Krishna Sathvik 1004286 | Pang Bang Yong 1004486 | Sarah Wong 1004659

Topic: Music Genre Classification

Genres in music are a way of generalizing different types of sounds and styles. It provides us with an understanding of sounds in music and how to differentiate songs from one other. This project aims to analyze the various features of audio clips from a variety of genres and produce a classifier able to extract the features of audio samples, such that it can categorize various sound samples into a specific musical genre with varying confidence.

Expected Inputs and Outputs

Inputs: 30 second audio data, class labels
Outputs: The genre of the audio sample

Data

GTZAN Dataset - Music Genre Classification

This dataset consists of 1000 audio samples of length 30 second across 10 different genres. The data is balanced across all classes, with 100 audio samples per class. The dataset also includes a collection of images representing the mel-frequency spectrogram representation of the audio data and two CSV files containing various features of the data such as mean and variance.

The genres included in this dataset are: Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, Rock

In this project, we will only be using the audio sample files and their corresponding class labels for training.

Architecture Draft

Given the temporal nature of audio data, we plan to use variants of a Recurrent Neural Network (RNN) such as the Long-Short Term Memory (LSTM) or Gated Recurrent Unit (GRU), or Transformer architectures.

There are various methods of feature extraction that may work well for our application of music audio data, such as calculating the Mel-frequency Cepstral Coefficients (MFCC), Gammatone-frequency Cepstral Coefficients (GFCC), or Wav2Vec2FeatureExtractor from the Wav2Vec2 Model, which learns powerful representations in audio data. We may also explore

spectral features of the audio data such as entropy, spectral flux and spread. Feature encoding may also be done through Wav2Vec2's Encoder. If necessary, we may also explore various audio augmentation methods such as time shifting or stretching, pitch shifting, time dilation, or adding noise in order to enable more meaningful feature extraction to improve the model accuracy.

If time permits, we may also compile a small validation dataset of labeled audio samples to perform inference testing on the model.

This project will be built with Pytorch and torchaudio.

Deliverables

1. Code used for data preprocessing and training of models
2. Trained model parameters
3. Code for model deployment and GUI tool
4. GUI application to use the model and sample audio files
5. Project Report